Language variety identification in Spanish tweets

Wolfgang Maier
Institute for Language and Information
University of Düsseldorf
Düsseldorf, Germany
maierw@hhu.de

Carlos Gómez-Rodríguez
Depto. de Computación
Universidade da Coruña
A Coruña, Spain
cgomezr@udc.es

Abstract

We study the problem of *language variant identification*, approximated by the problem of labeling tweets from Spanish speaking countries by the country from which they were posted. While this task is closely related to "pure" language identification, it comes with additional complications. We build a balanced collection of tweets and apply techniques from language modeling. A simplified version of the task is also solved by human test subjects, who are outperformed by the automatic classification. Our best automatic system achieves an overall F-score of 67.7% on 5-class classification.

1 Introduction

Spanish (or *castellano*), a descendant of Latin, is currently the language with the second largest number of native speakers after Mandarin Chinese, namely around 414 million people (Lewis et al., 2014). Spanish has a large number of regional varieties across Spain and the Americas (Lipski, 1994).[1] They diverge in spoken language and vocabulary and also, albeit to a lesser extent, in syntax. Between different American varieties of Spanish, there are important differences; however, the largest differences can be found between American and European ("Peninsular") Spanish.

Language identification, the task of automatically identifying the natural language used in a given text segment, is a relatively well understood problem (see Section 2). To our knowledge, however, there is little previous work on the identification of the varieties of a single language, such as the regional varieties of Spanish. This task is especially challenging because the differences between variants are subtle, making it difficult to discern between them. This is evidenced by the fact that humans that are native speakers of the varieties are often unable to solve the problem, particularly when given short, noisy text segments (which are the focus of this work) where the amount of available information is limited.

In this paper, we approximate the problem of language variety identification by the problem of classifying status messages from the microblogging service Twitter ("tweets") from Spanish speaking countries by the country from which they were sent. With the tweet, the location of the device from which the tweet was sent can be recorded (depending on the Twitter users' permission) and can then be retrieved from the metadata of the tweet. The tweet location information does not always correlate with the actual language variety used in the tweet: it is conceivable, e.g., that migrants do not use the prevalent language variety of the country in which they live, but rather their native variety. Nevertheless, Twitter can give a realistic picture of actual language use in a certain region, which, additionally, is closer to spoken than to standard written language. Eventually and more importantly, Twitter data is available from almost all Spanish speaking countries.

We proceed as follows. We build a balanced collection of tweets sent by Twitter users from five countries, namely Argentina, Chile, Colombia, Mexico, and Spain. Applying different methods, we perform an automatic classification between all countries. In order to obtain a more detailed view of the difficulty of our task, we also investigate human performance. For this purpose, we build a smaller sample of tweets from Argentina, Chile and Spain and have them classified by both our system and three native human evaluators. The results show that automatic classification outperforms human annotators. The best variant of our system, using a meta-classifier with voting,

[1] We are aware that there are natively Spanish-speaking communities elsewhere, such as on the Philippines, but we do not consider them in this study.

Language Technology for Closely Related Languages and Language Variants (LT4CloseLang), pages 25–35,
October 29, 2014, Doha, Qatar.

reaches an overall F-score of 67.72 on the five-class problem. On the two-class problem, human classification is outperformed by a large margin.

The remainder of this paper is structured as follows. In the following section, we present related work. Section 3 presents our data collection. Sections 4 and 5 present our classification methodology and the experiments. Section 7 discusses the results, and Section 8 concludes the article.

2 Related Work

Research on language identification has seen a variety of methods. A well established technique is the use of character n-gram models. Cavnar and Trenkle (1994) build n-gram frequency "profiles" for several languages and classify text by matching it to the profiles. Dunning (1994) uses language modeling. This technique is general and not limited to language identification; it has also been successfully employed in other areas, e.g., in authorship attribution (Kešelj et al., 2003) and author native language identification (Gyawali et al., 2013). Other language identification systems use non-textual methods, exploiting optical properties of text such as stroke geometry (Muir and Thomas, 2000), or using compression methods which rely on the assumption that natural languages differ by their entropy, and consequently by the rate to which they can be compressed (Teahan, 2000; Benedetto et al., 2002). Two newer approaches are Brown (2013), who uses character n-grams, and Řehůřek and Kolkus (2009), who treat "noisy" web text and therefore consider the particular influence of single words in discriminating between languages.

Language identification is harder the shorter the text segments whose language is to be identified (Baldwin and Lui, 2010). Especially due to the rise of Twitter, this particular problem has recently received attention. Several solutions have been proposed. Vatanen et al. (2010) compare character n-gram language models with elaborate smoothing techniques to the approach of Cavnar and Trenkle and the Google Language ID API, on the basis of different versions of the Universal Declaration of Human Rights. Other researchers work on Twitter. Bergsma et al. (2012) use language identification to create language specific tweet collections, thereby facilitating more high-quality results with supervised techniques. Lui and Baldwin (2014) review a wide range of off-the-shelf tools for Twitter language identification, and achieve their best results with a voting over three individual systems, one of them being `langid.py` (Lui and Baldwin, 2012). Carter et al. (2013) exploit particular characteristics of Twitter (such as user profile data and relations between Twitter users) to improve language identification on this genre. Bush (2014) successfully uses LZW compression for Twitter language identification.

Within the field of natural language processing, the problem of language variant identification has only begun to be studied very recently. Zampieri et al. (2013) have addressed the task for Spanish newspaper texts, using character and word n-gram models as well as POS and morphological information. Very recently, the Discriminating between Similar Languages (DSL) Shared Task (Zampieri et al., 2014) proposed the problem of identifying between pairs of similar languages and language variants on sentences from newspaper corpora, one of the pairs being Peninsular vs. Argentine Spanish. However, all these approaches are tailored to the standard language found in news sources, very different from the colloquial, noisy language of tweets, which presents distinct challenges for NLP (Derczynski et al., 2013; Vilares et al., 2013). Lui and Cook (2013) evaluate various approaches to classify documents into Australian, British and Canadian English, including a corpus of tweets, but we are not aware of any previous work on variant identification in Spanish tweets.

A review of research on Spanish varieties from a linguistics point of view is beyond the scope of this article. Recommended further literature in this area is Lipski (1994), Quesada Pacheco (2002) and Alvar (1996b; 1996a).

3 Data Collection

We first built a collection of tweets using the Twitter streaming API,[2] requesting all tweets sent within the geographic areas given by the coordinates -120°, -55° and -29°, 30° (roughly delimiting Latin America), as well as -10°, 35° and 3°, 46° (roughly delimiting Spain). The download ran from July 2 to July 4, 2014. In a second step, we sorted the tweets according to the respective countries.

Twitter is not used to the same extent in all countries where Spanish is spoken. In the time

[2] `https://dev.twitter.com/docs/api/streaming`

it took to collect 2,400 tweets from Bolivia, we could collect over 700,000 tweets from Argentina.[3] To ensure homogeneous conditions for our experiments, our final tweet collection comprises exactly 100,000 tweets from each of the five countries from which most tweets were collected, that is, Argentina, Chile, Colombia, Mexico, and Spain.

At this stage, we do not perform any cleanup or normalization operations such as, e.g., deleting forwarded tweets ("re-tweets"), deleting tweets which are sent by robots, or tweets not written in Spanish (some tweets use code switching, or are entirely written in a different language, mostly in English or in regional and minority languages that coexist with Spanish in the focus countries). Our reasoning behind this is that the tweet production in a certain country captures the variant of Spanish that is spoken.

We mark the start and end of single tweets by <s> and </s>, respectively. We use 80% of the tweets of each language for training, and 10% for development and testing, respectively. The data is split in a round-robin fashion, i.e., every ninth tweet is put into the development set and every tenth tweet is put in the test set, all other tweets are put in the training set.

In order to help with the interpretation of classification results, we investigate the distribution of tweet lengths on the development set, as shown in Figure 1. We see that in all countries, tweets tend to be either short, or take advantage of all available characters. Lengths around 100 to 110 characters are the rarest. The clearest further trend is that the tweets from Colombia and, especially, Argentina tend to be shorter than the tweets from the other countries.

4 Automatic Tweet Classification

The classification task we envisage is similar to the task of language identification in short text segments. We explore three methods that have been used before for that task, namely character n-gram frequency profiles (Cavnar and Trenkle, 1994; Vatanen et al., 2010), character n-gram language models (Vatanen et al., 2010), as well as LZW compression (Bush, 2014). Furthermore, we explore the usability of syllable-based language models. For all four approaches, we train models for binary classification for each class, i.e., five models that decide for each tweet if it belongs to a single class. As final label, we take the output of the one of the five classifiers that has the highest score.

We finally use a meta-classifier on the basis of voting. All methods are tested on the development set. For evaluation, we compute precision, recall and F_1 overall as well as for single classes.

Note that we decided to rely on the tweet text only. An exploration of the benefit of, e.g., directly exploiting Twitter-specific information (such as user mentions or hash tags) is out of the scope of this paper.

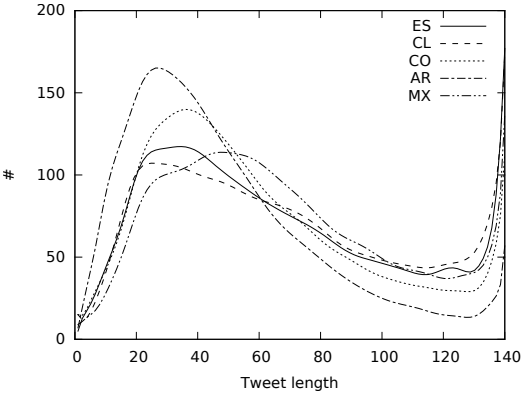

Figure 1: Tweet length distribution

	50	100	500	1k	10k
AR	31.68	29.72	43.93	31.77	18.42
CO	24.29	21.36	26.14	19.68	19.03
MX	31.86	28.97	32.58	30.28	22.27
ES	20.19	25.22	22.08	21.25	16.15
CL	22.95	29.74	35.67	26.01	16.69

Table 1: Results (F_1): n-gram frequency profiles (classes/profile sizes)

4.1 Character n-gram frequency profiles

We first investigate the n-gram frequency approach of Cavnar and Trenkle (1994). We use the well-known implementation `TextCat`.[4] The results for all classes with different profile sizes are shown in Table 1. Table 2 shows precision and recall for the best setting, a profile with a maximal size of 500 entries.

The results obtained with a profile size of 500

[3] We are aware that the Twitter API does not make all sent tweets available. However, we still assume that this huge difference reflects a variance in the number of Twitter users.

[4] As available from `http://odur.let.rug.nl/~vannoord/TextCat/`.

class	precision	recall	F_1
AR	32.60	67.33	43.93
CO	31.66	22.26	26.14
MX	51.52	23.82	32.58
ES	32.83	16.63	22.08
CL	31.96	40.36	35.67
overall	34.08	34.08	34.08

Table 2: Results: n-gram frequency profile with 500 n-grams

	AR	CO	MX	ES	CL
AR	6,733	949	384	610	1,324
CO	4,207	2,226	720	803	2,044
MX	2,547	1,342	2,382	1,051	2,678
ES	3,781	1,361	649	1,663	2,546
CL	3,384	1,153	488	939	4,036

Table 3: Confusion matrix (n-gram freq. profiles, 500 n-grams)

entries for Colombia align with the results for Spain and Mexico in that the precision is higher than the recall. The results for Chile align with those for Argentina with the recall being higher than the precision. For Mexico and Argentina the differences between recall and precision are particularly large (28 and 35 points, respectively). The confusion matrix in Table 3 reveals that tweets from all classes are likely to be mislabeled as coming from Argentina, while, on the other hand, Mexican tweets are mislabeled most frequently as coming from other countries.

Overall, the n-gram frequency profiles are not very good at our task, achieving an maximal overall F-score of only 34.08 with a profile size of 500 entries. However, this performance is still well above the 20.00 F-score we would obtain with a random baseline. Larger profile sizes deteriorate results: with 10,000 entries, we only have an overall F-score of 18.23. As observed before (Vatanen et al., 2010), the weak performance can most likely be attributed to the shortness of the tweets and the resulting lack of frequent n-grams that hinders a successful profile matching. While Vatanen et al. alleviate this problem to some extent, they have more success with character-level n-gram language models, the approach which we explore next.

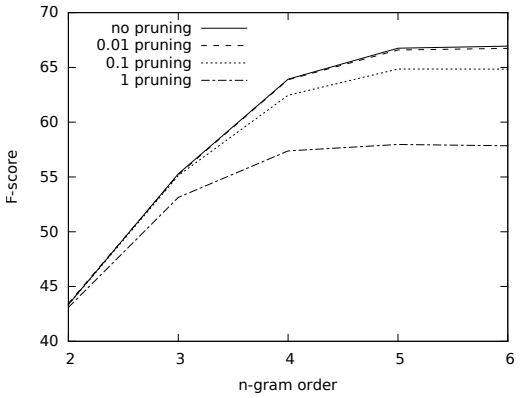

Figure 2: Character n-gram lm: Pruning vs. n-gram order

4.2 Character n-gram language models

We recur to n-gram language models as available in `variKN` (Siivola et al., 2007).[5] We run `variKN` with absolute discounting and the cross-product of four different pruning settings (no pruning, and thresholds 0.01, 0.1 and 1) and five different n-gram lengths (2 to 6).

Figure 2 contrasts the effect of different pruning settings with different n-gram lengths. While excessive pruning is detrimental to the result, slight pruning has barely any effect on the results, while reducing look-up time immensely. The order of the n-grams, however, does have an important influence. We confirm that also for this problem, we do not benefit from increasing it beyond $n = 6$, like Vatanen et al. (2010).

We now check if some countries are more difficult to identify than others and how they benefit from different n-gram orders. Figure 3 visualizes the corresponding results. Not all countries profit equally from longer n-grams. When comparing the 3- and 6-gram models without pruning, we see that the F_1 for Argentina is just 8 points higher, while the difference is more than 14 points for Mexico.

Table 4 shows all results including precision and recall for all classes, in the setting with 6-grams and no pruning. We can see that this approach works noticeably better than the frequency profiles, achieving an overall F-score of 66.96. The behavior of the classes is not uniform: Argentina shows the largest difference between precision and recall, and is furthermore the only class in which precision is higher than recall. Note also that in

[5] `https://github.com/vsiivola/variKN`

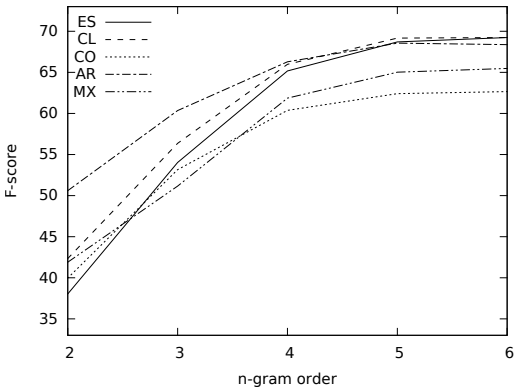

Figure 3: Character n-gram lm: Classes vs. n-gram order (no pruning)

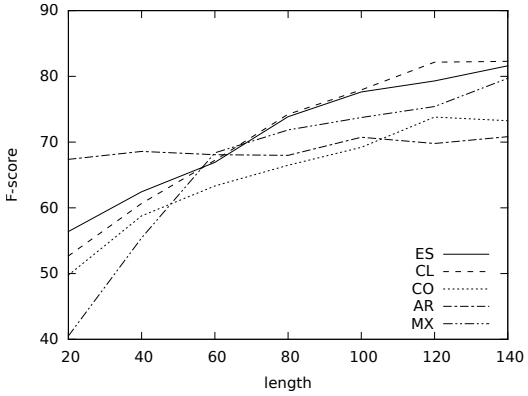

Figure 4: Character n-grams: Results (F_1) for tweet length intervals

class	precision	recall	F_1
AR	70.67	66.22	68.37
CO	62.56	62.77	62.66
MX	65.23	65.74	65.48
ES	68.75	69.36	69.06
CL	67.81	70.73	69.24
overall	66.96	66.96	66.96

Table 4: Results: 6-grams without pruning

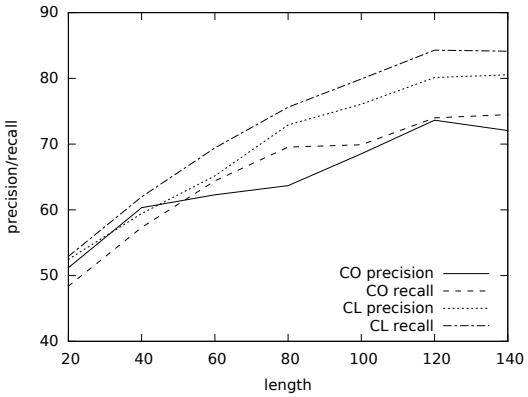

Figure 5: Character n-grams: Precision/recall for AR and CL

general, the differences between precision and recall are lower than for the n-gram frequency profile approach. The confusion matrix shown in Table 5 reveals that the Colombia class is the one with the highest confusion, particularly in combination with the Mexican class. This could indicate that those classes are more heterogeneous than the others, possibly showing more Twitter-specific noise, such as tweets consisting only of URLs, etc.

We finally investigate how tweet length influences classification performance in the 6-gram model. Figure 4 shows the F-scores for intervals of length 20 for all classes. The graph confirms that longer tweets are easier to classify. This correlates with findings from previous work. Over 82 points F_1 are achieved for tweets from Chile longer than 120 characters, while for those containing up to 20 characters, F_1 is almost 30 points lower. We investigate precision and recall separately. Figure 5 shows the corresponding curves for the best and worst performing classes, namely, CL and CO. For Chile, both precision and recall develop in parallel to the F-score (i.e., the longer the tweets, the higher the scores). For Colombia, the curves confirm that the low F_1 is rather due to a low precision than a low recall, particularly for tweets longer than 40 characters. This correlates with the counts in the confusion table (Tab. 5).

	AR	CO	MX	ES	CL
AR	6,622	1,036	702	740	900
CO	800	6,277	1,151	875	897
MX	509	1,237	6,574	847	833
ES	630	850	857	6,936	727
CL	809	634	794	690	7,073

Table 5: Confusion matrix (6-grams, no pruning)

4.3 Syllable n-gram language models

Since varieties of Spanish exhibit differences in vocabulary, we may think that models based on word n-grams can be more useful than character n-grams to discriminate between varieties. However, the larger diversity of word n-grams means that such models run into sparsity problems. An intermediate family of models can be built by us-

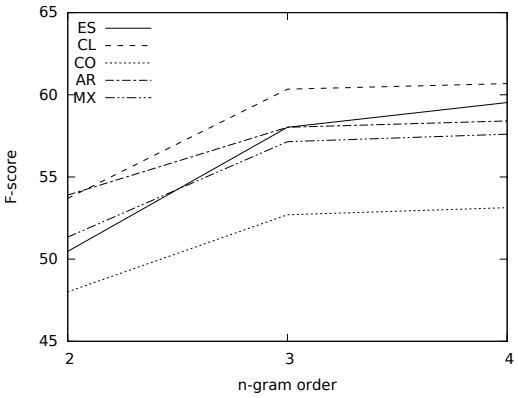

Figure 6: Syllable n-gram lm: pruning vs. n-gram order

ing syllable n-grams, taking advantage of the fact that Spanish variants do not differ in the criteria for syllabification of written words. Note that this property does not hold in general for the language identification problem, as different languages typically have different syllabification rules, which is a likely reason why syllable n-gram models have not been used for this problem.

To perform the splitting of Spanish words into syllables, we use the TIP syllabifier (Hernández-Figueroa et al., 2012), which applies an algorithm implementing the general syllabification rules described by the Royal Spanish Academy of Language and outlined in standard Spanish dictionaries and grammars. These rules are enough to correctly split the vast majority of Spanish words, excluding only a few corner cases related with word prefixes (Hernández-Figueroa et al., 2013). While accurate syllabification requires texts to be written correctly with accented characters, and this is often not the case in informal online environments (Vilares et al., 2014); we assume that this need not cause problems because the errors originated by unaccented words will follow a uniform pattern, producing a viable model for the purposes of classification.

We train n-gram language models with `variKN` as described in the last section, using absolute discounting. Due to the larger vocabulary size, we limit ourselves to 0.01 pruning, and to n-gram orders 2 to 4. Figure 6 shows the results (F_1) of all classes for the different n-gram orders, and Table 6 shows the results for all classes for the 4-gram language model.

As expected, shorter n-grams are more effective for syllable than for character language models.

class	precision	recall	F_1
AR	55.94	61.11	58.41
CO	53.23	53.03	53.13
MX	59.10	56.17	57.60
ES	62.35	56.96	59.53
CL	59.31	62.12	60.68
overall	57.88	57.88	57.88

Table 6: Results (F_1): Syllable 4-gram lm

For the Chilean tweets, e.g., the F-score for the 2-gram language model is around 11 points higher than for the character 2-gram language model. Furthermore, the performance seems to converge earlier, given that the results change only slightly when raising the n-gram order from 3 to 4. The overall F-score for the 4-gram language model is around 6 points lower than for character 4-grams. However, the behavior of the classes is similar: again, Mexico and Colombia have slightly lower results than the other classes.

4.4 Compression

We eventually test the applicability of compression-based classification using the approach of Bush (2014). As mentioned earlier, the assumption behind compression-based strategies for text categorization is that different text categories have a different entropy. Classification is possible because the effectivity of compression algorithms depends on the entropy of the data to be compressed (less entropy $\approx$ more compression).

A simple classification algorithm is Lempel-Ziv-Welch (LZW) (Welch, 1984). It is based on a dictionary which maps sequences of symbols to unique indices. Compression is achieved by replacing sequences of input symbols with the respective dictionary indices. More precisely, compression works as follows. First, the dictionary is initialized with the inventory of symbols (i.e., with all possible 1-grams). Then, until the input is fully consumed, we repeat the following steps. We search the dictionary for the longest sequence of symbols s that matches the current input, we output the dictionary entry for s, remove s from the input and add s followed by the next input symbol to the dictionary.

For our experiments, we use our own implementation of LZW. We first build LZW dictionaries by compressing our training sets as described

	1k	8k	25k	50k
AR	28.42	38.78	46.92	51.89
CO	19.81	28.27	32.81	36.05
MX	22.07	33.90	43.10	45.06
ES	22.08	29.48	35.15	38.61
CL	27.08	28.22	33.59	36.68

Table 7: Results (F_1): LZW without ties

class	precision	recall	F_1
AR	70.96	68.36	69.64
CO	62.44	64.22	63.32
MX	66.37	65.67	66.02
ES	70.10	69.64	69.87
CL	68.97	70.72	69.83
overall	67.72	67.72	67.72

Table 8: Results: Voting

above, using different limits on dictionary lengths. As symbol inventory, we use bytes, not unicode symbols. Then we use these dictionaries to compress all tweets from all test sets, skipping the initialization stage. The country assigned to each tweet is the one whose dictionary yields the highest compression. We run LZW with different maximal dictionary sizes.

The problem with the evaluation of the results is that the compression produced many ties, i.e., the compression of a single tweet with dictionaries from different languages resulted in identical compression rates. On the concatenated dev sets (50k tweets, i.e., 10k per country) with a maximal dictionary size of 1k, 8k, 25k and 50k entries, we got 14.867, 20,166, 22,031, and 23,652 ties, respectively. In 3,515 (7%), 4,839 (10%), 5,455 (11%) and 6,102 (12%) cases, respectively, the correct result was hidden in a tie. If we replace the labels of all tied instances with a new label TIE, we obtain the F-scores shown in Table 7. While they are higher than the scores for n-gram frequency profiles, they still lie well below the results for both syllable and character language models.

While previous literature mentions an ideal size limit on the dictionary of 8k entries (Bush, 2014), we obtain better results the larger the dictionaries. Note that already with a dictionary of size 1000, even without including the ties, we are above the 20.00 F-score of a random baseline. The high rate of ties constitutes a major problem of this approach, and remains even if we would find improvements to the approach (one possibility could be to use unicode characters instead of bytes for dictionary initialization). It cannot easily be alleviated, because if the compression rate is taken as the score, particularly the scores for short tweets are likely to coincide.

4.5 Voting

Voting is a simple meta-classifying technique which takes the output of different classifiers and decides based on a predefined method on one of them, thereby combining their strengths and leveling out their weaknesses. It has been successfully used to improve language identification on Twitter data by Lui and Baldwin (2014).

We utilize the character 5-gram and 6-gram language models without pruning, as well as the syllable 3-gram and 4-gram models. We decide as follows. All instances for which the output of the 5-gram model coincides with the output of at least one of the syllable models are labeled with the output of the 5-gram model. For all other instances, the output of the 6-gram model is used. The corresponding results for all classes are shown in Table 8.

We obtain a slightly higher F-score than for the 6-gram character language model (0.8 points). In other words, even though the 6-gram language model leads to the highest overall results among individual models, in some instances it is outperformed by the lower-order character language model and by the syllable language models, which have a lower overall score.

5 Human Tweet Classification

In order to get a better idea of the difficulty of the task of classifying tweets by the country of their authors, we have tweets classified by humans.

Generally, speakers of Spanish have limited contact with speakers of other varieties, simply due to geographical separation of varieties. We therefore recur to a simplified version of our task, in which the test subjects only have to distinguish their own variety from one other variety, i.e., perform a binary classification. We randomly draw two times 150 tweets from the Argentinian test and 150 tweets from the Chilean and Spanish test sets, respectively. We then build shuffled concatenations of the first 150 Argentinian and the Chilean tweets, as well as of the remaining 150 Argentinian and the Spanish tweets. Then we let three

data	subject	class	prec.	rec.	F_1
AR-ES	AR	AR	68.5	76.7	72.3
		ES	73.5	64.7	68.8
	ES	AR	71.5	62.0	66.4
		ES	66.5	75.3	70.6
	n-gram	AR	**92.3**	**87.3**	**89.7**
		ES	**88.0**	**92.7**	**90.3**
AR-CL	AR	AR	61.0	77.3	68.2
		CL	69.1	50.7	58.5
	CL	AR	70.0	70.0	70.0
		CL	70.0	70.0	70.0
	n-gram	AR	**93.4**	**84.7**	**88.8**
		CL	**86.0**	**94.0**	**89.8**

Table 9: Results: Human vs. automatic classification

data	subject	class	prec.	rec.	F_1
AR-ES	AR	AR	71.8	80.0	75.7
		ES	74.8	65.4	69.7
	ES	AR	74.6	62.9	68.2
		ES	65.1	76.3	70.2
	n-gram	AR	**93.2**	**88.6**	**90.8**
		ES	**88.1**	**92.9**	**90.4**
AR-CL	AR	AR	61.1	78.6	68.8
		CL	68.8	48.5	56.9
	CL	AR	73.0	71.4	72.2
		CL	71.2	72.8	72.0
	n-gram	AR	**95.3**	**87.1**	**91.0**
		CL	**87.8**	**95.6**	**91.5**

Table 10: Results: Human vs. automatic classification (filtered)

natives classify them. The test subjects are not given any other training data samples or similar resources before the task, and they are instructed not to look up on the Internet any information within the tweet that might reveal the country of its author (such as hyperlinks, user mentions or hash tags).

Table 9 shows the results, together with the results on the same task of the character 6-gram model without pruning. Note that with 300 test instances out of 20,000, there is a sampling error of $\pm$ 4.7% (confidence interval 95%). The results confirm our intuition in the light of the good performance achieved by the n-gram approach in the 5-class case: when reducing the classification problem from five classes to two, human classification performance is much below the performance of automatic classification, by between 17 and 31 F-score points. In terms of error rate, the human annotators made between 3 and 4 times more classification errors than the automatic system. One can observe a tendency among the human test subjects that more errors come from labeling too many tweets as coming from their native country than vice versa (cf. the recall values).

In order to better understand the large result difference, we ask the test subjects for the strategies they used to label tweets. They stated that the easiest tweets where those specifying a location ("*Estoy en Madrid*"), or referencing local named entities (TV programs, public figures, etc.). In case of absence of such information, other clues were used that tend to occur in only one variety. They include the use of different words (such as *enfadado* (Spain) vs. *enojado* (America) ("angry")), a different distribution of the same word (such as the filler *pues*), and different inflection, such as the second person plural verb forms, which in American Spanish, albeit sometimes not in Chile, is replaced by the identical third person plural forms (for the verb *hacer* ("do"), the peninsular form would be *hacéis* instead of *hacen*), and the personal pronoun *vos* ("you"), which is rarely used in Chile, and not used in Spain. To sum up, the test subjects generally relied on lexical cues on the surface, and were therefore bound to miss non-obvious information captured by the character n-gram model.

Since the test subjects also stated that some tweets were impossible to assign to a country because they contained only URLs, emoticons, or similar, in Table 10 we show a reevaluation of a second version of the two shuffled concatenated samples in which we remove all tweets which contain only emoticons, URLs, or numbers; tweets which are entirely written in a language other than Spanish; and tweets which are only two or one words long (i.e., tweets with zero or one spaces). For the AR-ES data, we remove 23 Spanish and 10 Argentinian tweets, while for the AR-CL data, we remove 10 Argentinian and 14 Chilean tweets.

As for the human classification on the AR/ES data, the results for Spain do not change much. For Argentina, there is an increase in performance (2 to 3 points). On the AR/CL data, there is a slight improvement on all sets except for the Chilean data classified.

As for the automatic classification, the filtering gives better result on all data sets. However,

	training	dev	test
AR	57,546 (71.9%)	7,174	7,196
CO	58,068 (72.6%)	7,249	7,289
MX	48,527 (60.7%)	6,117	6,061
ES	53,199 (66.5%)	6,699	6,657
CL	56,865 (71.1%)	6,998	7,071

Table 11: Data sizes (filtered by `langid.py`)

class	precision	recall	F_1
AR	70.32	66.09	68.14
CO	63.76	62.22	62.98
MX	61.52	61.11	61.31
ES	69.13	69.20	69.17
CL	67.12	73.29	70.07
overall	66.45	66.45	66.45

Table 12: Results: Filtered by `langid.py`

the difference between the F_1 of the filtered and unfiltered data is larger on the AR/CL data set. This can be explained with the fact that among the tweets removed from the AR/ES data set, there were more longer tweets (not written in Spanish) than among the tweets removed from the CL/AR data set, the longer tweets being easier to identify. Note that the filtering of tweets does not cause much change in the difference between human and automatic classification.

6 Language Filtering

As mentioned before, our data has not been cleaned up or normalized. In particular, the data set contains tweets written in languages other than Spanish. We have reasoned that those can be seen as belonging to the "natural" language production of a country. However, in order to see what impact they have on our classification results, we perform an additional experiment on a version of the data were we only include the tweets that the state-of-the-art language identifier `langid.py` labels Spanish (Lui and Baldwin, 2012).[6] Table 11 shows the sizes of all data sets after filtering. Note that many of the excluded tweets are in fact written in Spanish, but are very noisy, due to orthography, Twitter hash tags, etc. The next most frequent labels across all tweets is English (9%). Note that in the data from Spain, 2% of the tweets are labeled as Catalan, 1.2% as Galician, and only 0.3% as Basque.

Table 12 finally shows the classification results for character 6-gram language models without pruning.

The changes in F_1 are minor, i.e., below one point, except for the Mexican tweets, which lose around 4 points. The previous experiments have already indicated that the Mexican data set is the most heterogeneous one which also resulted in the largest number of tweets being filtered out. In general, we see that the character n-gram method seems to be relatively stable with respect to a different number of non-Spanish tweets in the data. More insight could be obtained by performing experiments with advanced methods of tweet normalization, such as those of Han and Baldwin (2011). We leave this for future work.

[6] `https://github.com/saffsd/langid.py`.

7 Discussion

Human classification of language varieties was judged by our test subjects to be considerably more difficult that differentiating between languages. Additionally, the test subjects were only able to differentiate between two classes. While the automatic classification results lie below the results which one would expect for language identification, n-gram classification still achieves good performance.

Our experiments touch on the more general question of how a language variety is defined. In order to take advantage of the metadata provided by Twitter, we had to restrict the classification problem to identifying varieties associated with countries were tweets were sent. In reality, the boundaries between variants are often blurred, and there can also be variance within the same country (e.g., the Spanish spoken in the southern Spanish region of Andalusia is different from that of Asturias, even if they both share features common to Peninsular Spanish and larger differences with American Spanish). However, it would be difficult to obtain a reliable corpus with this kind of fine-grained distinctions.

It is also worth noting that not all the classification criteria used by the human test subjects were purely linguistic – for example, a subject could guess a tweet as being from Chile by recognizing a mention to a Chilean city, public figure or TV show. Note that this factor intuitively seems to benefit humans – who have a wealth of knowledge about entities, events and trending topics from their country – over the automatic system. In spite

of this, automatic classification still vastly outperformed human classification, suggesting that the language models are capturing linguistic patterns that are not obvious to humans.

8 Conclusion

We have studied different approaches to the task of classifying tweets from Spanish-speaking countries according to the country from which they were sent. To the best of our knowledge, these are the first results for this problem. On the problem of assigning one of five classes (Argentina, Mexico, Chile, Colombia, Spain) to 10,000 tweets, the best performance, an overall F-score of 67.72, was obtained with a voting meta-classifier approach that recombines the results for four single classifiers, the 6-gram (66.96 F_1) and 5-gram (66.75 F_1) character-based language models, and the 4-gram (57.87 F_1) and 3-gram (57.24 F_1) syllable-based language models. For a simplified version of the problem that only required a decision between two classes (Argentina vs. Chile and Spain vs. Argentina), given a sample of 150 tweets from each class, human classification was outperformed by automatic classification by up to 31 points.

In future work, we want to investigate the effect of tweet normalization on our problem, and furthermore, how the techniques we have used can be applied to classify text from other social media sources, such as Facebook.

Acknowledgements

The first author has been funded by Deutsche Forschungsgemeinschaft (DFG). The second author has been partially funded by Ministerio de Economía y Competitividad/FEDER (Grant TIN2010-18552-C03-02) and by Xunta de Galicia (Grant CN2012/008).

References

Manuel Alvar, editor. 1996a. *Manual de dialectología hispánica. El español de América*. Ariel, Barcelona.

Manuel Alvar, editor. 1996b. *Manual de dialectología hispánica. El español de España*. Ariel, Barcelona.

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, CA.

Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. 2002. Language trees and zipping. *Physical Review Letters*, 88(4).

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 65–74, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ralph D. Brown. 2013. Selecting and weighting n-grams to identify 1100 languages. In Springer, editor, *Proceedings of the 16th International Conference on Text, Speech, and Dialogue*, volume 8082 of *LNCS*, pages 475–483, Pilsen, Czech Republic.

Brian O. Bush. 2014. Language identication of tweets using LZW compression. In *3rd Pacific Northwest Regional NLP Workshop: NW-NLP 2014*, Redmond, WA.

Simon Carter, Wouter Weerkamp, and Manos Tsagkias. 2013. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, NV.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 198–206. Association for Computational Linguistics.

Ted Dunning. 1994. Statistical identification of language. Technical Report MCCS-94-273, Computing Research Lab, New Mexico State University.

Binod Gyawali, Gabriela Ramirez, and Thamar Solorio. 2013. Native language identification: a simple n-gram based approach. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–231, Atlanta, Georgia, June. Association for Computational Linguistics.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA, June. Association for Computational Linguistics.

Zenón Hernández-Figeroa, Gustavo Rodríguez-Rodríguez, and Francisco J. Carreras-Riudavets. 2012. Separador de sílabas

del español - silabeador TIP. Available at http://tip.dis.ulpgc.es.

Zenón Hernández-Figueroa, Francisco J. Carreras-Riudavets, and Gustavo Rodríguez-Rodríguez. 2013. Automatic syllabification for Spanish using lemmatization and derivation to solve the prefix's prominence issue. *Expert Syst. Appl.*, 40(17):7122–7131.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of PACLING*, pages 255–264.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, editors. 2014. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, seventeenth edition edition. Online version: http://www.ethnologue.com.

John M. Lipski. 1994. *Latin American Spanish*. Longman, London.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea, July. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden.

Marco Lui and Paul Cook. 2013. Classifying english documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 5–15, Brisbane, Australia, December.

Douglas W. Muir and Timothy R. Thomas. 2000. Automatic language identification by stroke geometry analysis, May 16. US Patent 6,064,767.

Miguel Ángel Quesada Pacheco. 2002. *El Español de América*. Editorial Tecnológica de Costa Rica, Cartago, 2a edition.

Vesa Siivola, Teemu Hirsimäki, and Sami Virpioja. 2007. On growing and pruning kneser-ney smoothed n-gram models. *IEEE Transactions on Speech, Audio and Language Processing*, 15(5):1617–1624.

William J. Teahan. 2000. Text classification and segmentation using minimum cross-entropy. In *Proceedings of RIAO'00*, pages 943–961.

Tommi Vatanen, Jaakko J. Vyrynen, and Sami Virpioja. 2010. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2013. Supervised polarity classification of spanish tweets based on linguistic knowledge. In *Proceedings of 13th ACM Symposium on Document Engineering (DocEng 2013)*, pages 169–172, Florence, Italy.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2014. A syntactic approach for opinion mining on Spanish reviews. *Natural Language Engineering*, FirstView:1–25, 6.

Radim Řehůřek and Milan Kolkus. 2009. Language identification on the web: Extending the dictionary method. In *Proceedings of CICLing*, pages 357–368.

Terry A. Welch. 1984. A technique for high-performance data compression. *Computer*, 17(6):8–19, June.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and pos distribution for the identification of spanish varieties. In *Proceedings of TALN2013*, pages 580–587.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the dsl shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.